9П

УДК 004.934

В.В. Пилипенко, В.В. Робейко
Международный научно-учебный центр информационных технологий и систем,
г. Киев, Украина
valery_pylypenko@mail.ru, valya.robeiko@gmail.com

Автоматизированный стенограф украинской речи

В статье рассматривается автоматизированный стенограф для получения текста стенограммы из звукового файла на основе системы распознавания речи с участием оператора. Записанная фонограмма обрабатывается системой распознавания слитной речи многих дикторов из больших словарей (больше 10 тыс. слов). Оператор исправляет допущенные ошибки для получения текста, пригодного для дальнейшей работы. Он также вводит новые слова, не знакомые системе распознавания. На основе анализа ошибок и новых слов производится дообучение системы распознавания, что позволяет улучшать показатели надежности распознавания речи в процессе эксплуатации системы стенографирования.

Стенографирование широко используется для обработки и документирования материалов заседаний и совещаний различного уровня, для работы секретарей, журналистов и так далее. Компьютеры значительно расширили возможности и позволили увеличить гибкость применения систем стенографирования. На данный момент становится актуальным уменьшить долю ручного труда в таких системах. Для этого предлагается использовать автоматическое распознавание речи для превращения звука в текст.

Речь каждого человека сугубо индивидуальна. Поэтому перевести звук в текст по нажатию одной кнопки – задача довольно сложная для системы стенографирования. Такая система должна максимально упростить работу оператора и ускорить перевод звукового файла в текстовый, а также учесть все особенности речи диктора. Существует много программно-аппаратных комплексов автоматизированного стенографирования с различными возможностями, но даже самый простой позволяет увеличить скорость перевода звука в текст в несколько раз.

Автоматическое распознавание слитной речи многих дикторов из больших словарей значительно упрощает работу оператора, сводя её до исправления ошибок, допущенных системой стенографирования. Дообучение системы позволяет сокращать количество ошибок в процессе эксплуатации.

1. Автоматизированная VS автоматическая

Системы стенографирования можно условно разделить на три категории в зависимости от соотношения участия человека и компьютера в процессе создания стенограмм.

1. **Автоматические** (без участия человека в процессе распознавания речи).
2. **Автоматизированные** (компьютер распознает поток речи, человек участвует в этом процессе в той или иной степени).
3. **Стенографирование** при помощи компьютера (человек набирает текст, а компьютер используется как магнитофон и печатная машинка).

Разница между **автоматической** и **автоматизированной** системами заключается в надежности автоматического распознавания речи.

Опыт эксплуатации показывает, что первичная стенограмма, созданная человеком, содержит ошибки, которые исправляются в процессе редактирования набранного текста. В среднем количество ошибок достигает 5 на одну страницу текста, что составляет надежность 98 %, поскольку одна страница содержит приблизительно 2000 знаков или 250 слов. Таким образом, система стенографирования становится **автоматической** при надежности распознавания речи выше 98 %.

Такая надежность уже сегодня достижима для автоматического распознавания речи при некоторых ограничениях. При этом распознается речь только одного диктора. Для изолированно произносимых слов словарь достигает 15 тыс. слов, а для слитной речи такая надежность достигается при словаре в 1 тыс. слов.

Поэтому на настоящий момент актуальным является создание программ распознавания речи, свободных от таких ограничений. Для стенографирования необходимо достигнуть объемов словаря от 10 тыс. слов до нескольких миллионов. Количество задействованных дикторов от 100 до одной тысячи. При этом должна распознаваться слитная речи в реальном времени для современных компьютеров.

Автоматизированную систему стенографирования имеет смысл применять при надежности 80 % и выше. При этом оператору необходимо будет исправлять не более, чем каждое пятое слово в тексте, что можно делать при прослушивании звуковой дорожки в процессе ее воспроизведения.

2. Система распознавания слитной речи

В данной работе как базовая система используется инструментарий HTK [1] на основе скрытых Марковских моделей (СММ). Инструментарий HTK использовался для построения акустических и лингвистических моделей. Для распознавания речи был разработан программный комплекс, совместимый с акустическими и лингвистическими моделями HTK.

2.1. Пользовательский вид программы

Пользовательский вид программы стенографирования приведен на рис. 1. В верхнем окне схематически изображается осциллограмма звуковой дорожки с автоматически выделенными сегментами речи (фразами или синтагмами). Оператор выделяет нужный ему сегмент и прослушивает его. При этом он может просмотреть ответ распознавания, который можно исправить в случае ошибки. После редактирования ответ добавляется в стенограмму и автоматически происходит переход к следующему сегменту.

Пользователь имеет возможность перейти к нужному диктору, или прослушать необходимый сегмент стенограммы.

Распознавание производится автоматически в фоновом режиме работы программы. Все ошибки распознавания фиксируются и после того, как закончилось формирование стенограммы, происходит дообучение системы стенографирования. При этом в обучающую выборку добавляются новые слова и информация о новых дикторах. Таким образом, надежность распознавания повышается в процессе эксплуатации системы стенографирования.

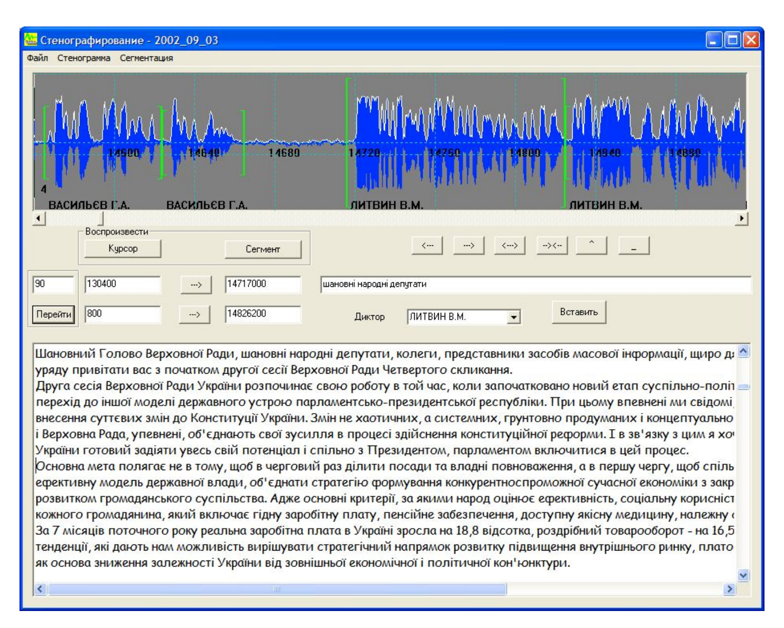


Рисунок 1 – Общий вид программы стенографирования

2.2. Предварительная обработка речевого сигнала

Речевой сигнал преобразуется в последовательность векторов признаков с интервалом анализа 25 мс и шагом анализа 10 мс. Вначале речевой сигнал фильтруется фильтром высоких частот с характеристикой $P(z) = 1 - 0{,}97z^{-1}$. Затем применяется окно Хэмминга и вычисляется быстрое преобразование Фурье. Спектральные коэффициенты усредняются с использованием 26 треугольных окон, расположенных в мел-шкале, и вычисляются 12 кепстральных коэффициентов.

Логарифм энергии добавляется в качестве 13-го коэффициента. Эти 13 коэффициентов расширяются до 39-мерного вектора параметров путем дописывания первой и второй разностей от коэффициентов, соседних по времени. Для учета влияния канала применяется вычитание среднего кепстра.

2.3. Акустическая модель

В качестве акустических моделей используются скрытые Марковские модели. 56 украинских контекстно-независимых фонем (включая фонему-паузу) моделируются тремя состояниями Марковской цепи без пропусков. Используется диагональный вид Гауссовских функций плотности вероятности.

Редко встречающиеся фонемы моделируются 64 смесями Гауссовских функций плотности вероятности, более часто встречающиеся фонемы моделируются большим числом смесей, наиболее часто встречающиеся фонемы используют 1024 смесей.

Словарь транскрипций создается автоматически из орфографического словаря с использованием контекстно-зависимых правил.

2.4. Многодикторная система

Распознавание речи независимо от диктора является задачей, не решенной до конца в распознавании речи. В [2] использовалась модель **кооперативного** распознавания, в которой при обучении смешивалась речь разных дикторов. При этом речь разных дикторов рассматривалась как разные реализации речи одного диктора.

Было показано, что надежность распознавания улучшалась не только для дикторов, входящих в **кооператив**, но также и для дикторов, незнакомых системе. Скорее всего, это связано с тем, что речь знакомых системе дикторов похожа на речь других дикторов.

Опыт применения такого подхода показал, что при использовании больше 100 дикторов в **кооперативе** надежность распознавания речи становится очень близкой к системе, независимой от диктора.

Методы работы с многими дикторами, заложенными в инструментарий НТК, такие как нормализация длины речевого тракта (Vocal Tract Length Normalisation) и адаптация модели при помощи линейного преобразования максимального правдоподобия (Maximum Likelihood Linear Regression), позволяют улучшить надежность распознавания речи для отдельных дикторов при условии, что каким-то независимым способом заранее определяется диктор. Предполагается использовать методы идентификации дикторов для автоматического определения говорящего.

3. Акустическое и текстовое наполнение

3.1. Обучающая выборка

Обучение производилось на выступлениях депутатов Верховной Рады Украины, записанных через телевизионную сеть. Парламентская речь характеризуется некоторыми особенностями:

1. Это спонтанная речь. Встречаются отдельные доклады, зачитываемые по подготовленному заранее тексту, однако мало дикторов в точности придерживаются этого текста.

2. Из-за ограничения во времени выступления многих дикторов произносятся в слишком быстром темпе.

3. Часто речь эмоционально окрашена.

4. В основном записи состоят из непрерывных выступлений дикторов, но в них встречаются реплики ведущего заседания или других депутатов.

5. Качество записи достаточно высокое, поскольку каждое депутатское место оснащено микрофоном.

Для обучения использовались записи, длиной в 99 тыс. секунд, в которых встретилось 211,224 слов. Всего было записано 208 дикторов. Дикторов, с длиной больше 300 с, оказалось 87. В табл. 1 приведено время записи для нескольких дикторов. Видно, что время записи сильно отличается для разных дикторов.

Таблица 1 – Время записи для нескольких дикторов в обучающей выборке

Диктор	Секунд
LIT	15805
POR	3715
SAH	2594
MOS	2490
MOR	1728

Диктор	Секунд
HAY	1620
SIM	1484
MAV	1305
ASA	1305
CHE	1140

Обучение производилось на предварительно размеченной выборке. Для этого запись выступления автоматически разбивалась на фразы из нескольких слов, ограниченные паузами больше 400 мс. Каждой фразе ставилась в соответствие метка в виде текста из стенограммы. Затем автоматически производилось преобразование текста в последовательность фонем в соответствии с контекстно-зависимыми правилами. Выборка, размеченная таким образом, использовалась для построения акустической модели.

3.2. Контрольная выборка

Распознавание производилось на выступлениях депутатов, записанных в отличные от обучающей выборки дни. Для распознавания использовались записи длиной в 30 тыс. секунд, в которых встретилось 68819 слов. Всего использовались записи 118 дикторов. Дикторов с длиной больше 300 с оказалось 37. Записи 36 дикторов не встретились в обучающей выборке. Таким образом, эти дикторы оказались неизвестными для системы распознавания. В табл. 2 приведено время записи нескольких дикторов.

Таблица 2 – Время записи для нескольких дикторов в контрольной выборке

Диктор	Секунд	Диктор	Секунд
LIT	4964	TES	567
SHL	1133	BAB	550
STO	887	TER	521
CHE	842	BAN	484
VAS	786	RUD	424

3.3. Текстовый материал

Словарь был составлен из текстов стенограмм заседаний Верховной Рады Украины. С официального сайта Верховной Рады были загружены все стенограммы заседаний, начиная с 1991 года, что составило больше 100 Мб текста. Текст был модифицирован для того, чтобы убрать служебную информацию из стенограмм (например, аплодисменты), записать числа в текстовом виде, а также отделить русский текст от украинского. Результирующий текст разделен на две части – первая содержит все тексты, кроме 2002 – 2003 годов, вторая содержит стенограммы 2002 – 2003 годов. Первая часть состоит из 14 629 111 слов, во второй содержится 409 244 слов.

Для первой части был составлен словарь из 156 108 слов и подсчитаны частоты встречаемости слов в этом словаре. Табл. 3 показывает долю текстов покрываемыми словами с определенными частотами. Видно, что весь словарь покрывает 99,6 % нового текста, не входящего в частотный словарь. Доля текста покрываемого словами с частотами выше 50 составляет больше 94 %. Для этого достаточно иметь словарь в 15 тыс. слов.

Таблица 3 – Доля текстов, входящих в частотные словари

Частота	Слов в частотном словаре	Слов этой частоты в тексте без 2002 – 2003 гг.	%	Слов этой частоты в тексте 2002 – 2003 гг.	%
1	156108	14629111	100,0	407608	99,6
2	98601	14571604	99,6	406563	99,3
3	78022	14530446	99,3	405753	99,1
5	58936	14465646	98,8	404293	98,7
10	40364	14343499	98,0	401544	98,1
50	**15609**	**13805357**	**94,3**	**388470**	**94,9**
100	10032	13415092	91,7	378873	92,5
200	6219	12878264	88,0	365510	89,3
300	4622	12488459	85,3	355606	86,8

Исследовалась надежность распознавания в зависимости от объема частотного словаря с использованием биграммной модели языка. Результаты представлены в табл. 4, из которой видно, что надежность незначительно увеличивается при увеличении размера словаря. Словаря объемом в 15 тыс. слов достаточно для распознавания речи с небольшим (2 %) уменьшением надежности от максимально возможного словаря.

Таблица 4 – Надежность распознавания для разных объемов частотного словаря

Объем словаря	64 000	50 000	30 000	20 000	15 000	10 000	5 000
Надежность распознавания, %	68,59	68,54	68,38	67,79	67,15	65,49	62,18

4. Биграммная модель языка

При распознавании речи использовалась биграммная модель речи, которая задавалась вероятностями появления пар слов. Поскольку в текстах, на которых вычислялись статистики, встретились далеко не все пары слов, возможные для данного словаря, то для аппроксимации *ненаблюдаемых* пар слов использовались так называемые **обратные** (back off) коэффициенты [1].

Биграммная модель языка позволила исправить много ошибок распознавания, которые не укладывались в модель языка. В табл. 5 показаны примеры исправления таких ошибок.

Таблица 5 – Примеры исправления ошибок распознавания при помощи биграммной модели языка

Было сказано	Свободный порядок слов	Биграммная грамматика
доброго ранку	до ***в в о*** ранку	доброго ранку
шановні народні депутати запрошені та гості Верховної Ради	шановні народі депутати запрошені та гості Верховної Ради	шановні народні депутати запрошені та гості Верховної Ради
прошу вас шановні колеги займати вас свої робочі місця	прошу ***в о з*** шановні колег **і й з е мати з в й й робоче** місця	прошу вас шановні колеги займати вас свої робочі місця
прошу займати робочі місця	прошу ***з е мав те й*** **робоче** ***й*** місця	прошу займати робочі місця
прошу підготуватися до реєстрації	прошу ***б й*** готуватися до реєстрації ***б***	прошу підготуватися до реєстрації

5. Модификация транскрипций

Для превращения орфографического текста в фонемный был разработан режим разбора орфографического текста и сформирован набор контекстно-зависимых правил, по которым орфографическое слово превращается в последовательность фонетических символов (путем преобразования одной последовательности символов в другую). Причем, генерируется сразу несколько вариантов транскрипции для случаев неоднозначностей, заданных в правилах.

Для всех дикторов был создан общий вариант транскрибирования. Кроме этого все дикторы были разделены на группы, для которых разработаны свои правила индивидуализированного транскрибирования, которые заменяют или дополняют основной вариант.

Результаты изучения речи многих дикторов свидетельствуют, что ни один из них не придерживается орфоэпических правил произношения в полном объеме. В первую очередь это касается запрещенных литературной нормой регрессивной ассимиляции за глухостью в паре фонем «звонкая + глухая» и оглушение согласных перед паузой. Дикторы с такими особенностями произношения были выделены в отдельную группу. Обработанный материал свидетельствует, что звонкие согласные в речи таких дикторов в позиции перед глухими оглушаются:

тобто → т о п т о

підтримати → п' і т т р И м а т и.

Случаи оглушения звонких согласных перед паузой встречаются у большинства дикторов: робив → р о б И ф.

Были выделены и многие другие характерные черты произношения разных дикторов: редуцирование окончаний некоторых слов (прилагательных, глаголов) в слитной речи (шановний → ш а н О в н и; доброго → д О б р о), «акание» (робити → р а б И т и), твёрдое произношение мягких согласных (синього → с И н о го) и др.

Такие тенденции моделируются путем изменения правил перехода от одних последовательностей символов к другим и расширением существующих правил.

Таблица 6 – Примеры модификации транскрипций слов

Слово	Обычная транскрипция	Модифицированная транскрипция
шановний	ш а н О в н и й	ш а н О в н и
коли	к о л И	к о л И к О л и к о л и
тільки	т' І л' к и	т' І л' к и т' і л' к и
при	п р И	п р И п р и
головуючий	г о л о в У й у ч и й	г о л о в У й у ч и
тобто	т О б т о	т о п т о
підтримати	п' і д т р И м а т и	п' і т т р И м а т и
народного	н а р О д н о г о	н а р О д н о
відповідно	в' і д п о в' І д н о	в' і т п о в' І д н о
робив	р о б И в	р о б И ф
доброго	д О б р о го	д О б р о
синього	с И н' о го	с И н о го
робити	р о б И т и	р а б И т и

В табл. 6 приведены примеры индивидуализированных транскрипций для нескольких слов. В основном здесь задействованы правила оглушения и редуцирования окончаний в словах. Для некоторых слов (служебных в том числе) задается несколько вариантов транскрипций – с ударением на разных слогах (если в языке возможны разные варианты прочтения таких слов) или вообще без ударения: коли → к о л И ; к О л и ; к о л и.

6. Результаты экспериментов по распознаванию слитной речи

Эксперименты проводились на описанной контрольной выборке в виде записей заседаний в течение одного дня. В табл. 7 приведены результаты распознавания записей для разных дней. Надежность распознавания сильно отличается в зависимости от того, какие дикторы попали в выборку. Например, в выборку DAY2 попал длинный доклад диктора SHL, который читал его скороговоркой. В четвертой колонке приведена надежность распознавания для индивидуальных модифицированных транскрипций. Применение индивидуальных транскрипций позволило улучшить надежность распознавания почти на один процент.

Таблица 7 – Надежность распознавания для разных выборок

Порция	Длина КВ, сек	Общие транскр. Надежность, %	Индивид. транскр. Надежность, %	Изменение, %
DAY1	1849	71,23	72,66	1,43
DAY2	5374	61,97	63,05	1,08
DAY3	10032	68,16	68,73	0,57
DAY3a	5990	68,69	69,75	1,06
DAY4	7260	76,66	77,13	0,47
Всего	**30505**	**69,28**	**70,06**	**0,78**

Табл. 8 представляет результаты распознавания порции DAY4 для некоторых дикторов, где также приведены длина обучающей выборки и темп произнесения для каждого диктора. Анализ результатов показывает, что в среднем указанные факторы (длина ОВ и темп речи) влияют на надежность распознавания. Надежность распознавания для отдельных дикторов сильно отличается – от 50 % до 90 %. Последняя колонка показывает изменение надежности распознавания речи для индивидуальных транскрипций. Видно, что для некоторых дикторов надежность ухудшается, для них следует применять другие индивидуальные транскрипции.

Таблица 8 – Надежность распознавания для нескольких дикторов

	Диктор	Длина ОВ, с	Длина КВ, с	Число слов в КВ	Темп, слов/с	Общие транскр., %	Индивид. транскр., %	Измене-ние, %
1	LIT	15805	2336	5721	2,45	79,85	80,56	0,71
2	POR	3715	411	853	2,08	80,30	80.54	0,24
3	MOR	1728	362	950	2,62	70,74	71,47	0,73
4	SIM	1484	125	255	2,04	80,00	80,78	0,78
5	MAT	1305	174	292	1,68	80,14	77,05	–3,09
6	KLU	998	107	209	1,95	86,60	89,0	2,40
7	KIN	585	223	417	1,87	64.27	66,43	2,16
8	ONI	483	100	209	2,09	79,90	80,38	0,48
9	MIS	195	148	312	2,11	69,87	69,23	–0,64
10	ZVA	25	101	205	2,03	80,0	80,49	0,49
11	GOL	0	379	790	2,08	78,48	78,35	–0,13
12	KAP	0	375	927	2,47	80,91	81,77	0,86
	…							
	Всего		**7260**	**16210**	**2,23**	**76,66**	**77,13**	**0,47**

Время распознавания для компьютера Pentium 2GHz составляет около 10 секунд для одной секунды речи. Применение алгоритмов ускорения принятия решений [3] позволит достичь реального времени распознавания речи.

Заключение

Статья описывает экспериментальную систему автоматизированного стенографирования. Показана возможность построения таких систем при условии повышения надежности распознавания речи до необходимых для практических применений показателей. Предложено использовать индивидуальную информацию о дикторах для улучшения надежности распознавания.

Литература

1. Young S., Evermann G., Kershaw D., Moore G., Odell J., Ollason D., Valtchev V., Woodland P. The HTK Book. – Cambridge University Engineering Department, 2002.
2. Винцюк Т.К. Анализ, распознавание и смысловая интерпретация речевых сигналов. – Киев: Наукова думка, 1987. – 264 с.
3. Пилипенко В.В. Распознавание дискретной и слитной речи из сверхбольших словарей на основе выборки информации из баз данных // Искусственный интеллект. – 2006. – № 3. – С. 548-557.

В.В. Пилипенко, В.В. Робейко

Автоматизований стенограф українського мовлення

У статті розглядається автоматизований стенограф для отримання тексту стенограми зі звукового файлу на основі системи розпізнавання мовлення за участі оператора. Записана фонограма опрацьовується системою розпізнавання мовлення кооперативу дикторів з великих словників (більше 10 тис. слів). Оператор виправляє допущені помилки для отримання тексту, придатного для подальшої обробки. Він також уводить нові слова, невідомі для системи розпізнавання. На основі аналізу помилок і нових слів проводиться донавчання системи розпізнавання, що дозволяє покращити показники надійності розпізнавання мовлення в процесі експлуатації системи стенографування.

Статья поступила в редакцию 22.07.2008.